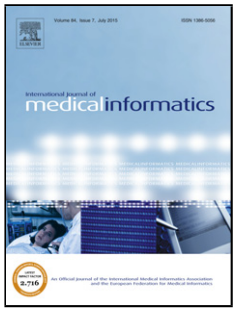
Journal Pre-proof



Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets

Mika Murtojärvi, Anni S. Halkola, Antti Airola, Teemu D. Laajala, Tuomas Mirtti, Tero Aittokallio, Tapio Pahikkala

PII: S1386-5056(18)31185-7

DOI: https://doi.org/10.1016/j.ijmedinf.2019.104014

Reference: IJB 104014

To appear in: *International Journal of Medical Informatics*

Received Date: 15 October 2018

Revised Date: 15 September 2019

Accepted Date: 15 October 2019

Please cite this article as: { doi: https://doi.org/

Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets

Mika Murtojärvi[a,*], Anni S. Halkola[b,c], Antti Airola[a], Teemu D. Laajala[b,c,d], Tuomas Mirtti[d,e,f], Tero Aittokallio[b,c,d], Tapio Pahikkala[a]

[a] *Department of Future Technologies, University of Turku, Turku, Finland*
[b] *Department of Mathematics and Statistics, University of Turku, Turku, Finland*
[c] *FICAN West Western Finland Cancer Centre*
[d] *Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland*
[e] *Department of Pathology, Medicum, University of Helsinki, Helsinki, Finland*
[f] *Department of Pathology, HUSLAB, Helsinki University Hospital, Helsinki, Finland*

Abstract

Introduction: Predictive survival modelling offers systematic tools for clinical decision-making and individualized tailoring of treatment strategies to improve patient outcomes while reducing overall healthcare costs. In 2015, a number of machine learning and statistical models were benchmarked in the DREAM 9.5 Prostate Cancer Challenge, based on open clinical trial data for metastatic castration resistant prostate cancer (mCRPC). However, applying these models into clinical practice poses a practical challenge due to the inclusion of a large number of model variables, some of which are not routinely monitored or are expensive to measure.

Objectives: To develop cost-specified variable selection algorithms for constructing cost-effective prognostic models of overall survival that still preserve sufficient model performance for clinical decision making.

Methods: Penalized Cox regression models were used for the survival prediction. For the variable selection, we implemented two algorithms: (i) LASSO regularization approach; and (ii) a greedy cost-specified variable selection algorithm. The models were compared in three cohorts of mCRPC patients from randomized clinical trials (RCT), as well as in a real-world cohort (RWC) of advanced prostate cancer patients treated at the Turku University Hospital. Hospital laboratory expenses were utilized as a reference for computing the costs of introducing new variables into the models.

Results: Compared to measuring the full set of clinical variables, economic costs could be reduced by half without a significant loss of model performance. The greedy algorithm outperformed the LASSO-based variable selection with the lowest tested budgets. The overall top performance was higher with the LASSO algorithm.

Conclusion: The cost-specified variable selection offers significant budget optimization capability for the real-world survival prediction without compromising the predictive power of the model.

Keywords: Cost optimization, survival prediction, prostate cancer, feature selection

1. Introduction

Prostate cancer is among the most commonly diagnosed types of cancer in men [1]. The survival time with the disease is highly dependent on its stage and grade at the time of diagnosis. Mortality due to low grade prostate cancer can be even lower than due to competing causes of death [2], whereas the median survival time of patients with newly diagnosed metastases is only 42 months [3]. Due to the aging population in many countries, the incidence and costs of prostate cancer are expected to increase significantly in the future [4, 5]. There is therefore a critical challenge to develop cost-effective procedures for prostate cancer management in order to minimize the burden on public health expenditure.

Several prognostic models have been developed for the survival prediction of patients with metastatic castration-resistant prostate cancer (mCRPC) [6, 7, 8, 9]. In the DREAM (Dialogue for Reverse Engineering Assessments and Methods) 9.5 mCRPC Challenge that was organized in 2015, 50 international teams developed competing models for overall survival prediction. The top performing model (ePCR; ensemble-based Penalized Cox Regression) significantly outperformed the competing models in independent validation [10]. Meta-analysis of the Challenge models supported the use of Cox regression model family coupled with regularization as the state of the art in survival prediction for patients with advanced prostate cancer [10, 11].

The prognostic models have conventionally been developed based on data from clinical trials where it is often possible to conduct a large number of laboratory tests. This may limit the usefulness of the models in clinical practice. For instance, the winning model of the DREAM competition takes as input 101 clinical variables ([10], supplementary appendix). Many of them are related to medical history and their extraction can be automated, but there are over 40 variables requiring laboratory tests or other rather expensive procedures such as imaging. Performing all required tests can incur a significant economic cost.

In order to develop cost effective prognostic models, one should favor approaches that require as few clinical variables as possible, while maintaining predictive performance. This task is closely related to *feature selection* (variable selection) that is well known in the field of machine learning [12]. Variable selection can serve a variety of purposes: models with a smaller number of variables are easier to interpret, more economic and faster to use, and may generalize better to new data. Varying approaches to variable selection have been explored when developing models for prostate cancer. In the Halabi model [9] 22 candidate variables were

*Corresponding author.

considered, of which 8 were included in the final model. In the ePCR model [10] some candidate variables were discarded due to being redundant, skewed or clinically insignificant according to expert evaluation. Such variables included clinical trial adverse effect variables, e.g. eye disorders. Model regularization was also used. Another model tested in the DREAM Challenge used survival forest and LASSO-based variable selection procedures [13]. However, none of the studies analyzed or proposed methods for minimizing the total economic cost of the variables required for applying the models in clinical settings.

The main contribution of this work is the development of cost-specified group-wise variable selection methods that are widely applicable to survival predictions based on patient hospital data generally available in clinical practice. The novel methods are applied here to patient cohorts of clinically significant prostate cancer, using real cost information from Finnish university hospitals. The results indicate that the approach enables significant saving of economic costs in a real-world settings, and yet high enough prediction accuracy for clinical applications. Four patient cohorts are included in the experiments: three originating from randomized clinical trials and one consisting of patients treated at the Turku University Hospital. Two variable selection methods are implemented and tested, one being a cost-specified greedy algorithm, and the other based on LASSO regularization. The methods are evaluated in two ways: by using cross-validation and by using different training and test cohorts. Experiments show that greedy selection gives better results when the allowed budget is so low that only a few variables are selected. On the other hand, the peak model performance with LASSO selection is higher in all cross-validation tests. Similar observations are made in tests with different training and test cohorts, although LASSO selection does not always outperform the greedy method even with a large budget.

2. Materials and methods

Penalized Cox regression [14] has previously been successfully applied to the survival prediction of patients with prostate cancer. For instance, the models by Halabi *et al.* [7, 9] and the top-performing model of the DREAM competition [10] were based on the Cox proportional hazards model coupled with regularization. Therefore, we also base our approach on penalized Cox regression. Three different types of penalization terms for the model coefficients are commonly used in order to prevent over-fitting the Cox model, L_1 (LASSO), L_2 (ridge regression), or their sum (elastic net) [15]. We only considered the former two possibilities in the variable selection.

2.1. Problem setting

In medical care, clinical laboratory tests are typically ordered as a package (group). For example, a standard blood test package at the Helsinki University Hospital laboratory contains nine measurements that are useful for characterizing the patient. This includes, for example, hemoglobin and counts of white and red blood cells and platelets. For the price of the laboratory test package one gets the results of all its measurements. Some measurements appear in more than one package.

Let the set of all available variables be F. A package of clinical tests is also called a *group of variables* G_i and can be specified by listing all its variables. Hence, $G_i \subseteq F$ for all $i \in \{1, 2, ..., n_g\}$, where n_g is the number of groups. To simplify the notation, individual measurements not belonging to any package are represented as groups containing one variable. It can then be assumed that all variables belong to some group, i.e. $\cup_{i=1}^{n_g} G_i = F$. On the other hand, a variable may belong to several groups. The price of a group G_i is denoted by c_i.

The variable groups that are included in a model can be specified by listing the corresponding group indices I. Thus, $I \subseteq \{1, 2, ..., n_g\}$. It is also allowed to include a group partially. Excluding variables of a group does not reduce cost but, due to the possibility of overfitting, it may improve model performance. The selected variables of group G_i are denoted by $s_i \subseteq G_i$ and the set of all selected variables by $S = \cup_{i \in I} s_i$. The total cost of the selected variables is $C = \sum_{i \in I} c_i$.

For given input data d, the performance of a model M is represented as a score function $score(M, d)$, with a higher score indicating a better model. When comparing models that only differ due to including different variables, one may also consider the score to be a function of the selected variables, $score(S, d)$. If a maximum budget B for making measurements is given, the problem is to maximize the performance $score(S, d)$ while respecting the budget constraint, i.e. $C \leq B$.

2.2. *Cost-specified* variable selection

The budget can be considered as a hard constraint on the set of selected variables, meaning that no violations of the constraint are allowed. It is difficult to enforce directly, because non-convexity and non-continuity makes the optimization NP-hard [16]. A popular way to select variables for Cox models is to introduce an L_1-norm constraint on the model, that can be considered as a convex and continuous approximation of the budget constraint. Using the Lagrange method, the L_1-norm constraint can be transformed to a so-called LASSO penalty function, a soft constraint whose effect is controlled by a penalty parameter. With a high enough amount of penalization, some of the model coefficients get a zero value. The corresponding variables can be removed. By varying the amount of penalization one can select different sets

of variables. Because the method does not take prices into account, price is computed afterwards using a heuristic. The heuristic starts with an empty variable set and adds variable groups sequentially. At each step the group G_i with the minimal cost per new variable is added. New variable means a variable that has not been already added and is in the target set S and the candidate variable group G_i. The process is continued until all variables of S have been included.

Using the LASSO penalty for variable selection has a side effect of also penalizing the coefficients of the useful variables that are selected. While this side effect is sometimes beneficial, as regularized models work better with noisy data, enforcing small budget constraints requires so strong regularization that it causes the Cox model to underfit (see the end of the supplementary appendix for a more detailed description and an example). Therefore, we propose an alternative technique that avoids the side effect, namely a greedy budget-constrained Cox regression (Greedy Cox) algorithm, that enforces the hard budget constraint directly. The algorithm can be seen as a variant of the Group-Wise Nested Forward Selection method proposed by Paclík *et al.* [17], though the selection criterion is not exactly the same, and their work did not concern Cox regression or survival analysis. The basic idea of the algorithm is to sequentially select the group of variables that gives, together with all variables that have been selected earlier, the best cross-validated prediction performance. This is locally optimal when the only allowed operation is the addition of a single group of variables. However, some variables of a group may not have a positive effect on prediction performance. The method is therefore further refined by selecting variables within the groups. This inner selection is similar to the group selection but operates on individual variables instead of groups, and the variables are restricted to those of the currently considered group. The variable selection process is stopped when there are no variable groups that fit within the remaining budget and improve prediction performance. We chose to utilize L_2 penalization in the models fitted in Greedy Cox, because when testing elastic net combining L_1 and L_2 in a similar setting [10], models close to using only L_2 were found to be optimal.

The pseudocode for selecting the next group and a subset of its variables is given in Algorithm 1. The algorithm uses the function *cv_score* to compute cross-validated (3-fold) estimates of model performance. To compute such an estimate, the function requires the $n \times d$ matrix X of all values of the clinical variables of the patients, the times and types of events $(\mathbf{y}, \delta)$ and the allowed variable set as an input. The entire variable selection process starts by initializing an empty set of selected variables. The remaining budget is set to the total budget because no groups have yet been selected. After the initialization Algorithm 1 is called repeatedly. After selecting each group the remaining budget and selected variables and groups are updated. When there are no variables available that fit within the budget and improve model performance,

the algorithm returns an empty set of variables, and the selection process is stopped.

Several alternatives are available for fitting the Cox models. Goeman [18] developed a method for LASSO models, and extensions to elastic net penalty were also outlined. The method uses both gradient descent and the Newton-Raphson algorithm. An algorithm by Simon *et al.* [19] is based on coordinatewise descent and was very fast in their tests. Wu [20] adapted least-angle regression to Cox models. Any method for fitting L_1- and L_2-penalized models is suitable for our purposes provided that the running time is not too high. The importance of running time results from both methods fitting a large number of models. In particular, in the greedy method every call of the function *cv_score* fits three models due to using 3-fold cross-validation.

Data: Values of variables X, Survival $(\mathbf{y}, \delta)$, Remaining budget b, Variable groups G, Group prices c, Selected variables S, Selected group indices I
Result: The index i of the best group to be added to I and the variables G'_i selected from G_i
$orig\_score := best\_score := cv\_score(X, (\mathbf{y}, \delta), S)$
$i, G'_i := null, \emptyset$
$I' := \{j | c_j \leq b \wedge j \notin I\}$
for $j \in I'$ **do** `// Iterate over groups`
  $F_{new} := G_j \setminus S$ `// Unselected variables of the current group`
  $group\_score, selected\_variables, selection\_finished := orig\_score, \emptyset, False$
  while *not selection_finished* **do** `// Find the best subset of variables`
    $S' := S \cup selected\_variables$
    $best\_variable\_score, best\_variable := group\_score, null$
    for $f \in F_{new}$ **do**
      $variable\_score := cv\_score(X, (\mathbf{y}, \delta), S' \cup \{f\})$
      if $variable\_score > best\_variable\_score$ **then**
        $best\_variable\_score, best\_variable := variable\_score, f$
    if $best\_variable \neq null$ **then**
      $group\_score, selected\_variables :=$
       $best\_variable\_score, selected\_variables \cup \{best\_variable\}$
    else
      $selection\_finished := True$
  if $group\_score > best\_score$ **then**
    $best\_score, i, G'_i := group\_score, j, selected\_variables$
return i, G'_i

Algorithm 1: Selecting the next group and its variables in the Greedy Cox algorithm. The remaining budget b is the total budget minus the price of the already selected groups I.

2.3. Model evaluation

Concordance index (C-index) [21] is selected as the primary measure of model performance, i.e. as the performance score function of Section 2.1. It is a measure of how well the order of modeled risks corresponds to the order of observed survival times. C-index has been commonly used in survival analysis [22, 23, 24, 25], including in the DREAM competition [10] (supplementary appendix). Although there

are several estimators for C-index with censored data [25], we limit to the version used in the DREAM competition: *survConcordance* function in the R package *survival*.

Cross-validation is used in model evaluation as follows. The final reported sets of variables are obtained by applying the selection algorithms on all data that are available for the studied patient cohort. In addition to the variable sets, estimates of model performance are required. For this purpose variable selection with five-fold cross-validation is repeated 50 times, giving a total of 250 sequences of variable sets. Each sequence contains all variable sets that were selected during a single run of a selection algorithm. The performance scores are computed during the cross-validation using the proper test sets. Finally, the results of the cross-validation are linked to the final variable sets (that were obtained using all data) by price. For a given variable set S this means that the last variable sets not exceeding the cost of S are selected from all 250 sequences and the corresponding 250 performance measures are averaged. Note that when using Greedy Cox, cross-validation is also used in the variable selection process: when considering a given cross-validation fold, the training data are further divided into cross-validation folds for the variable selection. The test fold is never included in the variable selection. This scheme is known as nested cross-validation [26].

In cross-validation tests the patient groups in the training and test sets tend to be highly similar because they are selected from the same patient cohort. This may lead to an optimistic bias in model evaluation. Therefore, further tests are done where the training and test cohorts originate from independent sources. We include four patient cohorts and consider all 12 pairs of training and test cohort where the cohorts are different.

2.4. Patient cohorts

Four patient cohorts were included, three of them originating from randomized clinical trials (RCT cohorts) included in the Prostate Cancer Challenge (PCC-DREAM), hosted by Project Data Sphere (PDS, https://www.projectdatasphere.org/), a broad-access research platform that collects and curates patient-level data from completed, phase III cancer clinical trials. The fourth group (real-world cohort, RWC) consists of patients treated at the Turku University Hospital according to the clinical recommendations. The patient registry data for the RWC cohort were provided by the Turku University Hospital Centre for Clinical Informatics and were processed as before [27]. A notification of the registry-based study design was made to the Office of the Data Protection Ombudsman according to the appropriate legislation, and the data gathering and analysis was performed with the permission of the hospital district (approval T287/2016). The patients of RWC were selected based on castration resistance [27].

The patient cohorts are summarized in Table 1. Only patients that received the standard treatment

(docetaxel and prednisone) were included in the three RCT cohorts. The ASCENT and MAINSAIL studies were terminated early due to the novel treatment not being beneficial in comparison to the standard treatment. The short follow-up times were reflected in mortality: less than 30 % of patients died during these trials, compared to over 70 % in the VENICE and RW cohorts.

The baseline patient characteristics for the RCT cohorts can be found in the respective publications [28, 29, 30]. In the RW cohort the median age of the patients was 76.3 years (first and third quartiles 70.1 and 82.6 years) in the beginning of the observation period, i.e. when their disease was first diagnosed as castration resistant. The dates of diagnosis ranged from April 2002 to October 2016. The median values of clinical variables were as follows: PSA 38.5 μg/l, HB 12.6 g/dl and alkaline phosphatase (ALP) 85 U/l. The median of the observed values of LDH was 193.5 U/l but the observation was missing for almost 90 % of the patients. Compared to the RCT cohorts, the patients in the RW cohort were older and their PSA and ALP values were lower. Information about metastases and the patient performance status (ECOG_C) were missing for most patients.

Table 1: Patient cohorts included in this study. The columns $T_{25\%/median/75\%}$ give the quartiles of days to event (death or censoring). *A data matrix collected and imputed during the DREAM competition was used. **Turku University Hospital.

Name	**Abbr.**	**Origin**	**Patients**	$\boldsymbol{T_{25\%}}$	$\boldsymbol{T_{median}}$	$\boldsymbol{T_{75\%}}$	$\boldsymbol{T_{max}}$	**% dead**
ASCENT	ASC	RCT[28] *	476	259	357	482	796	29.0
VENICE	VEN	RCT[29] *	598	392	643	902	1594	72.4
MAINSAIL	MAI	RCT[30] *	526	194	279	399	750	17.5
RW cohort	RWC	TYKS [27] **	581	128	330	702	4188	75.7

The survival curves for all four cohorts are shown in Figure 1. In the early follow-up survival is similar in all three RCT cohorts, whereas in the RW cohort early mortality is higher. This is reflected also in the median times to event in Table 1: although a similar number of deaths occurred in VEN and RWC, the median time to event is much higher in VEN. The short follow-up times in the ASCENT and MAINSAIL cohorts are also apparent. In the RW cohort, survival (or censoring) times are counted starting from the first instance mentioning castration resistance in the patient records. The PCA plot in Figure 1 indicates a difference in the baseline patient characteristics between RWC and the RCT cohorts.

As potential model variables we started with the 101 variables of the original ePCR model (Supplementary Data, Table 1). Prices of various clinical examinations were provided by the Helsinki University Hospital. Variables that can be automatically extracted from patient records, such as medical history, were assumed to be cost-free. There were 16 variables without a known cost, including information about metastases. Those variables were ignored, leaving 85 potential variables. The variable groups, their variables and prices are shown in Table 2.

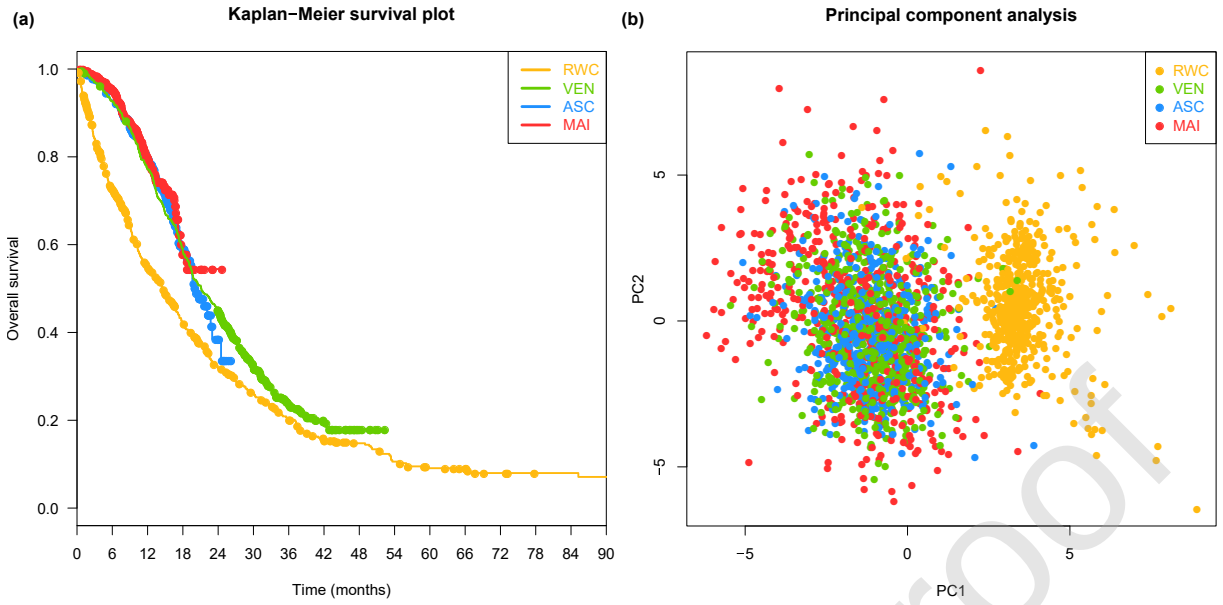


Figure 1: (a) Overall survival in the four cohorts. Circles indicate the censoring times. (b) PCA plot of the numeric variables of the four cohorts. The RCT cohorts appear to be similar to each other, but the RW cohort differs from them.

3. Results

The following questions were considered when conducting the cost vs. prediction performance trade-off evaluations:

1. Can variable set cost be reduced with little or no loss of model performance?
2. Which of the variable selection methods is optimal with different budgets?
3. Do the models give satisfactory results outside their training cohorts?

Cross-validation tests give answers to the first two questions. Tests where the training and test cohort originate from independent sources answer the last question while also providing further information about the usefulness of the estimated models. With the four patient cohorts there are 16 different pairs of training and test cohort: four where the cohorts are the same and 12 where they are different. All these pairs were considered by training a model on one cohort and evaluating it on another cohort. Cross-validation was used in cases where the training and test cohorts are the same. Several variables were missing for most of the patients in the RW cohort, including lactate dehydrogenase and aspartate aminotransferase, which have previously been identified as important predictive variables [10], resulting in differences in the selected variable sets between the RW and RCT cohorts.

Table 2: Variable groups and their prices. The prices are standardized so that the PSA measurement gets the reference cost of 100. The groups correspond to laboratory test packages available at the Helsinki University Hospital. Only variables considered for inclusion are shown. Non-abbreviated names of the variables are given in Supplementary Appendix, Table 1.

Group	**Model variables**	**Price**
B-PVKT	HB, HEMAT, RBC, WBC, PLT	40
B-PVK+Ne	Same as in B-PVKT except PLT	60
B-PVK+TKD	B-PVKT + LYMperLEU, MONOperLEU, NEU, POT, MONO, BASOperLEU, EOSperLEU, NEUperLEU	90
B-Hb	HB	50
cB-Het-Ion	NA.	100
Pt-GFReEPI	CCRC	20
P-LD	LDH	20
P-ASAT	AST	20
P-ALB	ALB	20
P-AFOS	ALP	20
P-PSA	PSA	100
P-Krea	CREAT	20
Pt-Krea-Cl	CREACL	70
B-Lymf	LYM	90
P-Pi	PHOS	20
P-Ca	CA	20
P-Alat	ALT	20
S-Prot	TPRO	20
B-PNH-La	WBC	4200
B-Eos	EOS	90
S-Testo	TESTO	330
P-Gluk	GLU	20
P-Mg	MG	20
P-Bil	TBILI	20
Free variables	Medical history, age, race/region, performance status (ECOG_C), medicines in use	0

3.1. Implementation

The variable selection algorithms described in Section 2.2 were implemented in Python language, version 3.5.2, and the *glmnet* package[1] was used for model fitting. Concordance indices were computed with the *lifelines* package. Variable sets were determined separately for the four data sets using both algorithms. In RWC, missing values were filled in using median imputation, which has previously been tested to give satisfactory results [27]. We also tested k-Nearest Neighbor (kNN) imputation before starting the variable selection tests but it did not improve model performance.

3.2. Test results

Figure 2 shows the C-index scores obtained in the four cohorts using the two variable selection algorithms with various budgets, both for the cross-validation and independent training-test cohort evaluations. There

[1] `https://web.stanford.edu/~hastie/glmnet_python/`

were clear differences between the cohorts in the achieved model performance. The order of survival times was predicted best in the RW cohort, with C-index up to 0.721, while in the VENICE cohort C-index was only 0.653. In the ASCENT and MAINSAIL cohorts the C-indices were up to 0.680 and 0.700, respectively. The reasons for the differences between the RCT cohorts are not known but the good results in the RW cohort may be explained by there not being stringent inclusion criteria in this cohort, making the patient population heterogeneous. In fact, PSA measurement alone allowed a model fitted to the RW cohort to achieve a C-index that was similar to the best tested models in the VENICE cohort.

3.2.1. Cross-validation results

In cross-validation tests the peak C-index of LASSO selection was always better than that of Greedy Cox but with low budgets Greedy Cox often achieved better results than LASSO. The results of LASSO selection also deteriorated with the greatest budgets while those of Greedy Cox remained stable. A possible reason is that in Greedy Cox the amount of model penalization is fixed while in LASSO selection high-cost variable sets are obtained by reducing the penalization.

With relatively low budgets slightly increasing the budget occasionally worsened model performance when using Greedy Cox. The reason is that the larger budget was spent on the first selected variable group, after which only cost-free variables could be selected. With a slightly lower budget, a less expensive variable group had to be selected first, leaving enough budget to select another inexpensive but non-free group.

Table 3 shows the variables selected for the RW cohort by the two algorithms and the achieved C-indices. An unlimited budget was used in Greedy Cox. The table shows the variables selected in each step of the algorithm. The results do not fully correspond to what one would get by limiting the budget, but it was verified that the differences are minor for the budgets shown in Table 3. The variable sets selected by the two methods are similar during the initial steps, after which the selection paths diverge. Of previously known prognostic variables lactate dehydrogenase (LDH) [9, 10] was not included, likely because it was missing for over 80 % of patients in RWC.

Table 3: Variables selected in the RW cohort in one representative run of the two algorithms. Variables selected after achieving peak performance score are not shown. An unlimited budget was used in Greedy Cox. Prices are standardized so that PSA measurement gets a reference cost of 100.

Added (LASSO)	**Cost**	**C-index**	**Added (Greedy)**	**Cost**	**C-index**
PSA	100	0.644	PSA	100	0.656
HB	140	0.675	HB, WBC, LYMperLEU, NEUperLEU	190	0.703

ALP	160	0.693	ALP	210	0.714
AGEGRP2	160	0.693	LYMPHAD.	210	0.714
NA.	260	0.711	MHRENAL	210	0.714
LYMPHAD., CEREBACC	260	0.711	CHF	210	0.714
CREAT	280	0.713	MHGASTRO	210	0.714
CA	300	0.716	MG	230	0.713
PROSTATECTOMY	300	0.716	CA	250	0.714
MHRENAL	300	0.716	MHRESP	250	0.714
MHRESP	300	0.716	CREAT	270	0.714
COPD	300	0.716	LDH	290	0.714
HEMAT	300	0.716	MHBLOOD	290	0.714
PHOS	320	0.719	MHCARD	290	0.714
AST	340	0.720	MHINJURY	290	0.714
BILAT. ORCHID.	340	0.720	CCRC	310	0.713
POT	430	0.721			
CHF, MHGASTRO	430	0.721			
RBC	430	0.721			
WBC	430	0.721			
HMG_COA_REDUCT	430	0.721			
ANALGESICS	430	0.721			
ACE_INHIBITORS	430	0.721			

3.2.2. Tests with different training and test cohorts

When different cohorts were used for model fitting and evaluation, Greedy Cox still often outperformed LASSO selection with low budgets. The peak performance of LASSO was again in many cases better than that of Greedy Cox, but there were also cases where Greedy Cox outperformed LASSO or their scores were very similar. Figure 2 also gives a coarse indication of how much model performance is lost when the training and test sets are not subsets of the same cohort, which is the more typical use case in practice. The cross-validated C-index in the MAINSAIL cohort with LASSO selection is clearly higher (0.70) than what is achieved by models fitted to the other cohorts (0.65-0.68). Greedy Cox applied on ASCENT does achieve a high C-index (slightly over 0.70) with a very low budget. However, in a real predictive setting

one would not know the C-index that will be achieved in the new cohort and the budget would have to be chosen based on, for example, the cross-validation results. Then, a greater budget would be chosen and the resulting model would perform much worse when evaluated in the MAINSAIL cohort, C-index well below 0.65. Except for the poor fit to MAINSAIL, a model fitted to the VENICE cohort performs well in the other cohorts. Models fitted to the RW cohort have somewhat similar predictive performance as those fitted to VENICE. Models fitted to MAINSAIL and ASCENT are not able to predict the survival times of the patients well for the RW and VENICE cohorts. The worse performance of models fitted to ASCENT or MAINSAIL compared to VENICE and RW may be related to the fact that there were relatively few deaths in the former two trials. In the RW cohort many observations are missing and, as noted before (Figure 1), the baseline patient characteristics differ between the RCT and RW cohorts.

4. Discussion

We developed tools for constructing survival models that incorporate both the prognostic value and real-life clinical cost of the available variables. The tools were applied to cohorts of prostate cancer patients, and considerable cost savings were possible. In particular, maximal prediction performance was obtained with variable sets whose total cost was 2-4 times the cost of PSA measurement. Additional variables had a negative effect on prediction performance when using LASSO penalization because the amount of penalization had to be reduced to include more variables. The number of variables in the best models was usually 10-15 and at most 19 when the full candidate set contained 85 variables with a known cost.

The variables selected by the two algorithms and for different cohorts (Supplementary Appendix, Section 4.2) were surprisingly dissimilar. Some of the differences are explained by data availability. For instance, lactate dehydrogenase (LDH) was selected in all RCT cohorts when the budget was sufficient for achieving maximal model performance, but in the RW cohort it was not available for most patients. However, even in the RCT cohorts LDH was the only variable that was always selected. When ranking the variables in terms of how often they were selected in the four cohorts by the two algorithms (8 test cases), LDH, hemoglobin (HB), alkaline phosphatase (ALP), history of congestive heart failure (CHF) and PSA ranked highest. Except for LDH and CHF, they were selected in the RW cohort by both algorithms. They were also selected in the RCT cohorts in 2-3 of 6 test cases. Apart from CHF these variables were also identified as important in the DREAM competition [10] (supplementary appendix). On the other hand, several variables that were important in the competition were rarely selected by the algorithms considered in this work: aspartate aminotransferase (AST), red blood cell count (RBC), albumin (ALB) and patient performance

status (ECOG_C). AST was selected in the RW cohort by both algorithms and in MAINSAIL cohort by Greedy Cox. ALB was selected in the MAINSAIL cohort by both algorithms but not in any other cohort. ECOG_C was selected only in the VENICE cohort by LASSO. RBC was not selected in any cohort by either algorithm.

In cross-validation tests greedy selection outperformed LASSO with low budgets but when a larger budget was allowed, LASSO was better. A similar observation has been made earlier in a different application domain [31] and has been explained by too much penalization when selecting only a few variables using LASSO [32]. When the training and test cohorts were different, the results were mixed between the variable selection methods. Overall, Greedy Cox was better when only a few variables were included in the models while LASSO was very competitive with larger budgets.

Further developments on both the greedy approach and penalized models are possible. For instance, in the greedy selection it might be beneficial to remove variables that have become redundant as a result of adding other variables. The penalized approach considered in this work, LASSO, does not take the groups of variables and their prices into account; the cost of a variable set was computed after the selection process. With larger budgets this gave rather good results, but incorporating the cost in the model as an additional penalization term might further improve the results, especially with lower budgets.

Limitations of the study

The study considered mortality due to all causes. Therefore the variables selected by the algorithms may contain variables that are not related specifically to prostate cancer. Even in the initial set of variables there was a scarcity of actual biomarkers of prostate cancer. While more comprehensive patient information would be preferable, the lack of biomarkers beyond PSA corresponds to what is available from the current clinical management of prostate cancer patients.

Metastasis status was excluded from variable selection due to missing price information and questionable availability in the RW data set. While this is a significant omission, one may note that in the 2015 DREAM competition information about any particular metastasis site was included in the models of less than 20 % of the teams, although the winning model did include several locations of metastases (see [27], Supplementary appendix). Unlike in the RCTs dealing with mCRPC, in the RW cohort there can be patients without any metastases, making the availability of this information potentially more important. Cost-specified variable selection should be repeated when the prices of additional variables become available. The proposed approach and the selected variables warrant further studies in other patient cohorts, both in prostate cancer and other

related cancer types, to confirm their prostate cancer-specificity, and applicability of the approach to other cancer types.

Authors' contributions

MM implemented and tested the variable selection algorithms and wrote most of the manuscript. ASH performed all tests where the ePCR package was involved and produced most of the figures. AA had a significant role in writing the manuscript and gave advice on developing the algorithms and conducting the tests. TDL gave advice on the best ways to use the ePCR package and provided feedback on the manuscript, as well as editing it himself. TM provided clinical expertise and participated in data curation and critically evaluated the manuscript and suggested changes. TA and TP supervised the study and participated in the writing process.

Acknowledgments

The authors thank Anna Hammais and Arho Virkki (Turku University Hospital, Centre for Clinical Informatics) for extracting and harmonizing the real-world patient data used in the study. A notification of the registry-based study design was made to the Office of the Data Protection Ombudsman according to the appropriate legislation, and the data gathering and analysis was performed with the study permission of Varsinais-Suomen sairaanhoitopiirin kuntayhtymä (approval T287/2016).

Funding

This work was supported by the Academy of Finland (grant 289903 to AA, grants 311273 and 313266 to TP, grants 295504 and 310507 to TA and grant 304667 to TM). TDL received funding from Finnish Cultural Foundation & Drug Research Doctoral Programme (DRDP). TA received funding from Cancer Society of Finland, Sigrid Juselius Foundation. ASH received funding from the University of Turku Doctoral Programme in Mathematics and Computer Sciences (MATTI). TM received funding from Cancer Society of Finland and Finnish Medical Foundation and research funding of Hospital District of Helsinki and Uusimaa.

Declarations of interest

None.

Summary table

What was already known on the topic

- Several models are available for the survival prediction of patients with prostate cancer
- Variable selection methods have been used when developing the models
- Variable set cost has not been an explicit goal in variable selection

What this study added to our knowledge

- Two methods were developed for selecting cost-effective sets of variables in survival prediction
- The methods offer significant cost reduction potential in all tested cohorts with minor or no loss of model performance
- The results of variable selection are sensitive to differences in algorithms and data sets
- With very low budgets Greedy Cox tends to produce better models than LASSO selection
- With large budgets LASSO selection is very competitive

References

[1] N. Mottet, J. Bellmunt, M. Bolla, E. Briers, M. G. Cumberbatch, M. De Santis, N. Fossati, T. Gross, A. M. Henry, S. Joniau, et al., Eau-estro-siog guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent, European urology 71 (4) (2017) 618–629.

[2] S. E. Eggener, P. T. Scardino, P. C. Walsh, M. Han, A. W. Partin, B. J. Trock, Z. Feng, D. P. Wood, J. A. Eastham, O. Yossepowitch, D. M. Rabah, M. W. Kattan, C. Yu, E. A. Klein, A. J. Stephenson, Predicting 15-year prostate cancer specific mortality after radical prostatectomy, The Journal of Urology 185 (3) (2011) 869 – 875.

[3] N. D. James, M. R. Spears, N. W. Clarke, D. P. Dearnaley, J. S. D. Bono, J. Gale, J. Hetherington, P. J. Hoskin, R. J. Jones, R. Laing, J. F. Lester, D. McLaren, C. C. Parker, M. K. Parmar, A. W. Ritchie, J. M. Russell, R. T. Strebel, G. N. Thalmann, M. D. Mason, M. R. Sydes, Survival with newly diagnosed metastatic prostate cancer in the "docetaxel era": Data from 917 patients in the control arm of the stampede trial (mrc pr08, cruk/06/019), European Urology 67 (6) (2015) 1028 – 1038.

[4] B. D. Smith, G. L. Smith, A. Hurria, G. N. Hortobagyi, T. A. Buchholz, Future of cancer incidence in the united states: Burdens upon an aging, changing nation, Journal of Clinical Oncology 27 (17) (2009) 2758–2765.

[5] C. G. Roehrborn, L. K. Black, The economic burden of prostate cancer, BJU International 108 (6) (2011) 806–813.

[6] O. Smaletz, H. I. Scher, E. J. Small, D. A. Verbel, A. McMillan, K. Regan, W. K. Kelly, M. W. Kattan, Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration, Journal of Clinical Oncology 20 (19) (2002) 3972–3982.

[7] S. Halabi, E. J. Small, P. W. Kantoff, M. W. Kattan, E. B. Kaplan, N. A. Dawson, E. G. Levine, B. A. Blumenstein, N. J. Vogelzang, Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer, Journal of Clinical Oncology 21 (7) (2003) 1232–1237.

[8] A. J. Armstrong, E. S. Garrett-Mayer, Y.-C. O. Yang, R. de Wit, I. F. Tannock, M. Eisenberger, A contemporary prognostic nomogram for men with hormone-refractory metastatic prostate cancer: a tax327 study analysis, Clinical Cancer Research 13 (21) (2007) 6396–6403.

[9] S. Halabi, C.-Y. Lin, W. K. Kelly, K. S. Fizazi, J. W. Moul, E. B. Kaplan, M. J. Morris, E. J. Small, Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer, Journal of Clinical Oncology 32 (7) (2014) 671.

[10] J. Guinney, T. Wang, T. D. Laajala, K. K. Winner, J. C. Bare, E. C. Neto, S. A. Khan, G. Peddinti, A. Airola, T. Pahikkala, et al., Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data, The Lancet Oncology 18 (1) (2017) 132–142.

[11] R. Meier, S. Graw, J. Usset, R. Raghavan, J. Dai, P. Chalise, S. Ellis, B. Fridley, D. Koestler, An ensemble-based cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mcrpc) patients [version 1; referees: 1 approved, 2 approved with reservations], F1000Research 5 (2677). doi:10.12688/f1000research.8226.1.
[12] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[13] S. Wengel Mogensen, A. H. Petersen, A. Buchardt, N. Hansen, Survival prognosis and variable selection: A case study for metastatic castrate resistant prostate cancer patients [version 1; referees: 2 approved], F1000Research 5 (2680). doi:10.12688/f1000research.8427.1.
[14] R. Tibshirani, The lasso method for variable selection in the cox model, Statistics in Medicine 16 (4) (1997) 385–395.
[15] M. Y. Park, T. Hastie, L1-regularization path algorithm for generalized linear models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (4) (2007) 659–677.
[16] S. Shalev-Shwartz, N. Srebro, T. Zhang, Trading accuracy for sparsity in optimization problems with sparsity constraints, SIAM Journal on Optimization 20 (6) (2010) 2807–2832. doi:10.1137/090759574.
[17] P. Paclík, R. P. W. Duin, G. M. P. van Kempen, R. Kohlus, On feature selection with measurement cost and grouped features, in: T. Caelli, A. Amin, R. P. W. Duin, D. de Ridder, M. Kamel (Eds.), Structural, Syntactic, and Statistical Pattern Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 461–469.
[18] J. J. Goeman, L1 penalized estimation in the cox proportional hazards model, Biometrical Journal 52 (1) (2010) 70–84.
[19] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for cox's proportional hazards model via coordinate descent, Journal of Statistical Software, Articles 39 (5) (2011) 1–13. doi:10.18637/jss.v039.i05.
[20] Y. Wu, Elastic net for cox's proportional hazards model with a solution path algorithm, Statistica Sinica 22 (1) (2012) 271–294.
[21] J. Harrell, Frank E., R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the Yield of Medical Tests, JAMA 247 (18) (1982) 2543–2546.
[22] M. W. Kattan, M. S. Karpeh, M. Mazumdar, M. F. Brennan, Postoperative nomogram for disease-specific survival after an r0 resection for gastric carcinoma, Journal of Clinical Oncology 21 (19) (2003) 3647–3650.
[23] W. G. Wierda, S. O'Brien, X. Wang, S. Faderl, A. Ferrajoli, K.-A. Do, J. Cortes, D. Thomas, G. Garcia-Manero, C. Koller, M. Beran, F. Giles, F. Ravandi, S. Lerner, H. Kantarjian, M. Keating, Prognostic nomogram and index for overall survival in previously untreated patients with chronic lymphocytic leukemia, Blood 109 (11) (2007) 4679–4685.
[24] B. Groot Koerkamp, J. K. Wiggers, M. Gonen, A. Doussot, P. J. Allen, M. G. H. Besselink, L. H. Blumgart, O. R. C. Busch, M. I. D'Angelica, R. P. DeMatteo, D. J. Gouma, T. P. Kingham, T. M. van Gulik, W. R. Jarnagin, Survival after resection of perihilar cholangiocarcinoma—development and external validation of a prognostic nomogram, Annals of Oncology 26 (9) (2015) 1930–1935.
[25] A. R. Brentnall, J. Cuzick, Use of the concordance index for predictors of censored survival data, Statistical Methods in Medical Research 27 (8) (2018) 2359–2373.
[26] D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, Journal of Cheminformatics 6 (10) (2014) .
[27] T. D. Laajala, M. Murtojärvi, A. Virkki, T. Aittokallio, ePCR: an R-package for survival and time-to-event prediction in advanced prostate cancer, applied to real-world patient cohorts, Bioinformatics (2018) bty477doi:10.1093/bioinformatics/bty477.
[28] H. I. Scher, X. Jia, K. Chi, R. de Wit, W. R. Berry, P. Albers, B. Henick, D. Waterhouse, D. J. Ruether, P. J. Rosen, A. A. Meluch, L. T. Nordquist, P. M. Venner, A. Heidenreich, L. Chu, G. Heller, Randomized, open-label phase iii trial of docetaxel plus high-dose calcitriol versus docetaxel plus prednisone for patients with castration-resistant prostate cancer, Journal of Clinical Oncology 29 (16) (2011) 2191–2198.
[29] I. F. Tannock, K. Fizazi, S. Ivanov, C. T. Karlsson, A. Fléchon, I. Skoneczna, F. Orlandi, G. Gravis, V. Matveev, S. Bavbek, et al., Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (venice): a phase 3, double-blind randomised trial, The lancet oncology 14 (8) (2013) 760–768.
[30] D. P. Petrylak, N. J. Vogelzang, N. Budnik, P. J. Wiechno, C. N. Sternberg, K. Doner, J. Bellmunt, J. M. Burke, M. O. de Olza, A. Choudhury, et al., Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (mainsail): a randomised, double-blind, placebo-controlled phase 3 trial, The Lancet Oncology 16 (4) (2015) 417–425.
[31] P. Naula, A. Airola, T. Salakoski, T. Pahikkala, Multi-label learning under feature extraction budgets, Pattern Recognition Letters 40 (2014) 56 – 65.
[32] T. Zhang, Adaptive forward-backward greedy algorithm for sparse learning with linear models, Advances in Neural Information Processing Systems 21 (2009) 1921–1928.

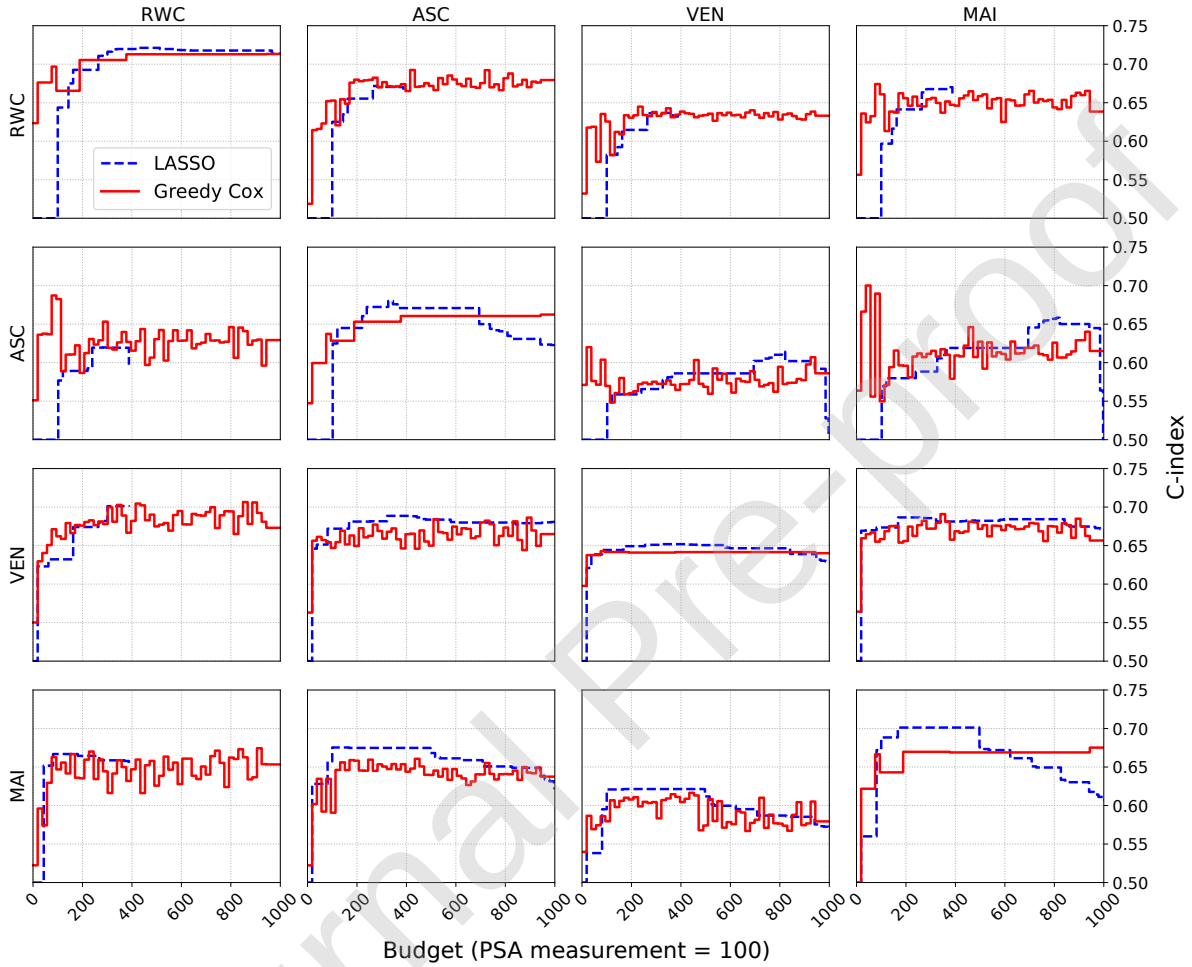


Figure 2: C-index scores achieved by the two variable selection algorithms in four cohorts. The rows correspond to the training cohorts and the columns to the test cohorts. The plots on the diagonal show cross-validation results where the training and test sets are non-overlapping subsets of the same cohort of patients. 50 budgets were considered when selecting variables using Greedy Cox, except in the cross-validation tests where only 8 budgets were tested due to high execution time. In tests involving the RW cohort, only variables that were available for at least 60 % of the patients were included as potential model variables. In other cases all variables with a known price were included in the analysis.